Chemometrics and Intelligent Laboratory Systems 142 (2015) 18–27

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Process monitoring based on factor analysis: Probabilistic analysis of monitoring statistics in presence of both complete and incomplete measurements

Zhonggai Zhao [a,*], Qinghua Li [a], Biao Huang [b], Fei Liu [a], Zhiqiang Ge [c]

[a] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China
[b] Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB T6G 2G6, Canada
[c] State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 4 June 2014
Received in revised form 21 August 2014
Accepted 23 December 2014
Available online 2 January 2015

Keywords:
Process monitoring
Factor analysis
Incomplete measurement
Probabilistic analysis

ABSTRACT

In generic process monitoring approaches based on probabilistic latent variable models, such as probabilistic principal component analysis (PPCA) or factor analysis (FA) model, the online score and residual are characterized by probability distributions. However, only their expectations are involved in the calculations of monitoring statistics, square prediction error (*SPE*) and Hotelling T^2, which ignore the information of their variances and may result in missed fault alarms. Based on the FA model, this paper investigates the probabilistic uncertainties of monitoring statistics arising from both inherent nature and missing measurements of the process data. The proposed method derives the distributions of both the online factor and residual at each sampling instant, and then transforms generic monitoring statistics into general quadratic forms. As a result, novel monitoring statistics are developed based on the probabilistic uncertainties of the generic statistics. In addition, the proposed monitoring statistics are extended to the case of incomplete measurements, in which the conditional distributions of the online measurement, factor and residual are computed and used to construct the statistics for process monitoring. Simulation examples illustrate the feasibility of the proposed method and demonstrate its effectiveness.

1. Introduction

Safe and efficient operation of industrial processes is always of interest to researchers as well as practitioners. Multivariate statistical process monitoring (MSPM) is an important technique to achieve this objective [1,2]. In MSPM, a statistical model is typically developed to capture the relation among a set of process variables under the normal operation condition. Two monitoring statistics, *SPE* and Hotelling's T^2, are usually developed for process monitoring. *SPE* is used to evaluate variation of residuals, and T^2 is an indicator for variation of factors or scores [3].

The principal component analysis (PCA) model is a well-known latent variable model [4]. In the PCA-based process monitoring, the online measurement is substituted into the PCA model to obtain the online residual and score, and then the monitoring statistic *SPE* is constructed by the Euclidian norm of residual and T^2 by the Mahalanobis norm of score [5]. Based on the assumption that process variables follow multivariate Gaussian distribution under normal operation condition [6], the PCA-based monitoring approach can achieve fairly good performance, thus it has been widely used in industrial processes [7,8]. However, it is difficult to extend the single PCA model to a mixture of such models for modeling the process whose operating conditions cannot be sufficiently captured by a Gaussian distribution [5,9,10]. Furthermore, since different measuring units are involved in constructions of *SPE* and T^2, it is also difficult to combine them to obtain consistent monitoring results [5].

Probabilistic PCA (PPCA) provides a fully probabilistic formulation of a Gaussian latent variable model [11]. PPCA was introduced for process monitoring by Kim [5], in which *SPE* and T^2were built respectively to monitor residuals and scores. Due to the same measurement units (Mahalanobis distance) that have been used in the two monitoring statistics, they can be easily unified to achieve a consistent monitoring result. On the other hand, it is also straightforward to extend the PPCA model to its mixture form [9], known as mixture PPCA models, which can be adopted for monitoring industrial processes subject to non-Gaussianity and multimodality [12].

However, a critical limitation of the PPCA model is that the noise levels are assumed to be identical for different process variables [11]. In fact, the noise may have different levels in different directions. If the different noise levels are assumed, the factor analysis (FA) model is obtained [13], which is a more general probabilistic latent variable model. FA has also been introduced into process monitoring and has shown good performance [14,15].

According to the FA or PPCA algorithm, both residuals and factors are probabilistic, however, the generic monitoring statistics *SPE* and T^2 are constructed only by the expectations of residuals and factors and their

* Corresponding author.
E-mail address: gaizihao@gmail.com (Z. Zhao).

http://dx.doi.org/10.1016/j.chemolab.2014.12.010
0169-7439/

variances are ignored, i.e. the useful uncertainty information is discarded, which may lead to misleading results.

On the other hand, missing measurements are commonly encountered in practice. How to deal with missing data is important for process monitoring, which can guarantee both the implementation of monitoring and the performance of monitoring. Walczak and Massart reviewed how to develop the PCA model with incomplete measurements [18,19]. Nelson and his coworkers analyzed the impact of missing measurements on the PCA and partial least square (PLS)-based monitoring, and evaluated the monitoring performance in the presence of missing measurements through the analysis of uncertainties arising from missing data [20,21]. For fault identification, Chen et al. made an assumption that a number of variables were missing and then derived the accompanying probabilistic certainties, then the variables resulting in larger probabilistic certainty were regarded as the faulty ones [22,23]. In the process monitoring based on the probabilistic PCA (PPCA) and PPCA mixture model, the contributions of each variable are calculated also by the missing variable approach [24]. To date, however, in the probabilistic model-based process monitoring approaches, how to develop monitoring statistics in the presence of missing measurements have not been reported.

Motivated by no use of the uncertainties of residuals and factors in constructions of monitoring statistics based on the FA model, this paper introduces the variance of residuals and factors into monitoring statistics and expresses monitoring statistics in probabilistic forms, and then derives their uncertain ranges by calculating their conditional distributions, then the novel statistics are developed by taking into account the uncertain information. Similarly, missing measurements will also result in uncertainties of monitoring statistics. This paper extends the uncertainty analysis of monitoring statistics to process monitoring in the presence of missing measurements, and re-builds monitoring statistics by considering the uncertainties arising from both the inherent nature and missing measurements. The remainder of this paper is organized as follows. The FA model and its monitoring scheme are briefly introduced in Section 2. Section 3 analyzes the uncertainty characteristics of *SPE* and T^2, based on which novel statistics are developed. In Section 4, the missing measurements are considered in the implementation of the FA-based process monitoring. The application of the proposed approach is illustrated in Section 5. Section 6 draws conclusions.

2. FA-based process monitoring

2.1. Nomenclature

Lower case bold letters or Greek and Roman letters are column vectors, bold capital ones are matrices and italic ones are scalars. Consider a normalized dataset $X = [x_1, x_2, ..., x_N]$ sampled under the normal operation condition, where N is the number of samples. The vector or matrix with superscript m and o indicates a vector or matrix with rows corresponding to missing variables and observable measurements, respectively. x_i represents the ith sample, and x is the measurement for any sample.

Without loss of generality, the ith sample with missing variables is assumed as $x_i = [x_{i,o}, x_{i,m}]$. According to the missing and observable variables, the variance of measurements $\mathbf{C}$ is organized into

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{oo} & \mathbf{C}_{om} \\ \mathbf{C}_{mo} & \mathbf{C}_{mm} \end{bmatrix}. \tag{1}$$

$\mathbf{W} = [w_1, w_2, \cdots, w_q] \in \Re^{k \times q}$ is the loading matrix, and k is generally larger than q. The bar on the vector denotes the expectation of the variable.

2.2. Maximum likelihood FA model

In the FA model, the k-dimensional normalized measurement x is generated from a q-dimensional score t with the residual or noise ϵ as

$$\mathbf{x} = \mathbf{Wt} + \epsilon. \tag{2}$$

Usually, the score and the noise are assumed to follow $G(\mathbf{0},\mathbf{I})$ and $\epsilon \sim G(\mathbf{0}, \boldsymbol{\Psi})$, respectively, where $\Psi = diag(\lambda_1, \lambda_2, \cdots, \lambda_q)$. The marginal distribution of measurements is represented by $\mathbf{x} \sim G(\mathbf{0}, \mathbf{WW}^T + \Psi)$, and the conditional distribution of measurement $p(\mathbf{x}|\mathbf{t})$ is derived as $G(\mathbf{W}t, \Psi)$. If $\lambda_1 = \lambda_2 = \cdots = \lambda_q$, the PPCA model is obtained. Therefore, the PPCA model is a special formulation of the FA model, and the PCA model is a special formulation of the PPCA model subjected to $\lambda_1 = \lambda_2 = \cdots = \lambda_q \approx 0$ [11].

According to Eq. (2), given the number of factors, the FA model is determined by $\mathbf{W}$ and Ψ. Maximum likelihood FA algorithm estimates these parameters by maximizing the likelihood of measurements [13,16]. However, $\mathbf{W}$ and Ψ are related to $\mathbf{t}$ but $\mathbf{t}$ is unknown, which makes the estimation difficult and there is no closed-form solution for the estimation. Treating $\mathbf{t}$ as missing data and the complete dataset as $(\mathbf{x},\mathbf{t})$, the expectation maximum (EM) algorithm can be employed to achieve the maximization of the likelihood. The EM algorithm is a useful tool to deal with the parameter estimation with missing measurements [25,26]. Writing the log-likelihood of complete-data as $\pounds_C = \sum_{n=1}^{N} \ln(p(\mathbf{x}_n, \mathbf{t}_n))$, the EM algorithm is accomplished through the following two iterative steps [13]:

E-step Fix the parameter and take the expectation of $\pounds_C$ with respect to the distribution $p(\mathbf{t}|x, W, \Psi)$. Omitting the independent terms of the model parameters, the expectation is expressed as

$$\begin{aligned} Q &= E\left[-\frac{N}{2}\ln|\Psi| - \sum_{n=1}^{N}\left(\frac{1}{2}\mathbf{x}_n^T\Psi^{-1}\mathbf{x}_n - \mathbf{x}_n^T\Psi^{-1}\mathbf{W}\mathbf{t}_n + \frac{1}{2}\mathbf{t}_n^T\mathbf{W}^T\Psi^{-1}\mathbf{W}\mathbf{t}_n\right)\right] \\ &= -\frac{N}{2}\ln|\Psi| - \sum_{n=1}^{N}\left(\frac{1}{2}\mathbf{x}_n^T\Psi^{-1}\mathbf{x}_n - \mathbf{x}_n^T\Psi^{-1}\mathbf{W}E[\mathbf{t}_n|\mathbf{x}_n] \right. \\ &\quad \left. + \frac{1}{2}tr\left[\mathbf{t}_n^T\Psi^{-1}E\left[\mathbf{t}_n\mathbf{t}_n^T|\mathbf{x}_n\right]\right]\right) \end{aligned} \tag{3}$$

where tr is the trace operator, $E(\mathbf{t}|\mathbf{x})$ and $E(\mathbf{tt}^T|\mathbf{x})$ are the expectations of $\mathbf{t}$ and $\mathbf{tt}^T$ conditioned on $\mathbf{x}$ and are calculated as

$$E(\mathbf{t}|\mathbf{x}) = \mathbf{W}^T C^{-1}\mathbf{x} \tag{4}$$

$$E\left(\mathbf{tt}^T|\mathbf{x}\right) = \mathbf{I} - \mathbf{W}^T\mathbf{C}^{-1}\mathbf{W} + \mathbf{W}^T\mathbf{C}^{-1}\mathbf{xx}^T\mathbf{C}^{-1}\mathbf{W} \tag{5}$$

where the variance of measurements $\mathbf{C} = \mathbf{WW}^T + \Psi$.

M-step Maximizing the expectation of log-likelihood by computing the derivation of log-likelihood with respect to these parameters and letting the derivation equal to 0, the parameter estimation is obtained as

$$\mathbf{W}^{new} = \left(\sum_{n=1}^{N}\mathbf{x}_n E[\mathbf{t}_n|\mathbf{x}_n]^T\right)\left(\sum_{n=1}^{N}E[\mathbf{t}_n\mathbf{t}_n^T|\mathbf{x}_n]\right)^{-1} \tag{6}$$

$$\Psi^{new} = \frac{1}{N}diag\left(\sum_{n=1}^{N}\left(\mathbf{x}_n\mathbf{x}_n^T - \mathbf{W}^{new}E[\mathbf{t}_n|\mathbf{x}_n]\mathbf{x}_n^T\right)\right). \tag{7}$$

The iteration of Eqs. (4), (5), (6) and (7) in sequence constitutes the EM algorithm to develop the FA model.

2.3. Process monitoring based on FA model

After the FA model has been built using the normal operation data, monitoring statistics should be constructed for on-line process